WP5: Task 51 (Decide on test data for scenario 1.3)

Authors:

Eva Toller (National Archives of Sweden, RA)
……

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	**Public**	**P**
C	**Confidential, only for members of the consortium and the Commission Services**	

Revision History

Revision	Date	Author	Organisation	Description
0.1	20130409	Eva Toller	RA	First draft
0.2	20130527	Eva Toller	RA	Documentation of the selected test data.

1 TEST DATA FOR SCENARIO 1.3

1.1 SCENARIO DESCRIPTION

"A little museum in Malta has a historical library and a digitised personal archive collection. The museum has staff of only 9 and only voluntary IT support. The director of the museum is aware of the need to organise digital preservation for the digitised documents, but is not sure how to do it. He receives periodically offers for long-term storage of digital content, but finds it difficult to select or to make a decision. He has practically no IT competence to rely on for decision-making, but is convinced that the decision should be forward-looking and accommodate the needs of the museum for the next 5 years."

General comment: if this scenario is reused in Proof of Concept #2, we could try to find a *real* organisation that has this problem (although it does not have to be a museum).

1.2 SUGGESTED TEST DATA

"Digitised documents": The same digital **DjVu** images that are used in Scenario 2.2 can also be used for Scenario 1.3 (see **DCH-RP_WP5_Scen-2-2_ID-66.doc**).

Scenario 1.3 is not dependent on using digitised data exclusively, so to make it a bit more realistic, the following data sets may be added:

- The data set from Scenario 2.4 (the "Linnéjubiléet" web site (Linné Anniversary), see **DCH-RP_WP5_Scen-2-4_ID-68.pdf**).
- A database extract called "Filmregistret" (Film/Movie index), records containing information about audited and censored films (directors, actors, cuts et c). It consists of structured text files, exported from a simple database like MS Access or FileMaker Pro. It does not have any security restrictions.
- Some spreadsheets from a government agency committee about Swedish constitutional law.

Also very small organisations tend to have at least a web site and some administrative records, so even if the contents of the website and the records are wrong, they should be possible to use in essence. Note that even for the DjVu images, it may be hard to find "appropriate" contents types.

1.3 PREPARATORY WORK

For the DjVu images and the web site, see **DCH-RP_WP5_Scen-2-2_ID-66.doc** and **DCH-RP_WP5_Scen-2-4_ID-68.pdf**, respectively.

General:

- Make all the necessary preparations and agreements with SweGrid/SweStore.

For “Filmregistret”:

- Obtain permission from the Electronic Archives section (“ElArk”, with Christina Olsson as their representative) at RA to use the records outside the Preservation Net (“Bevarandenätet”), although still within RAs internal net – or, if needs must, on a computer not connected to the Internet.
- Obtain permission from the legal owner of the information (“informationsägare”, with Karin Åström-Iko as their representative) to use the records outside RA – that is, to use them in the SweGrid/SweStore infrastructures.

1.4 DEPENDENCIES

The following chronological dependencies should be observed:

- Before any practical work (except for administrative preparations) can start, permission from the Electronic Archives section must be obtained.
- Before any data can be uploaded and processed in the SweGrid/SweStore infrastructures, permission from the legal owner of the information must be obtained.

1.5 DESCRIPTION OF SELECTED TEST DATA

1.5.1 Filmregistret (the Film Records Collection)

The data set “Filmregistret” is a records collection extracted from a relational database system. The records are not about films *per se*, but of the censorship activities that have been performed for some films scenes that contain violence or are offensive in other ways. The actual cuts are *not* included in this data set.

There are four separate files in “Filmregistret”:

Filmregistret.csv: contains general and administrative information about the films and the censoring process (including the name of the censors and the technicians). Other examples are:

- the category of film (for example, documentary)
- the distributor of the film
- the country of production
- the reason for the censoring
- age limit for seeing the film
- the number of cuts
- the length of the film before and after the cuts

All in all, there are 109 columns in *Filmregistret.csv*. The number of record instances (rows) is 59785.

Regissoer.csv: contains the names of the directors and IDs for the censorships for that director's films. There are only these 2 columns in this file. The number of record instances (rows) is 30405.

Skaadespelare.csv: contains the names of the actors and IDs for the censorships for that actor's films. There are only these 2 columns in this file. The number of record instances (rows) is 139301.

Klipp.csv: contains detailed information about the cuts that have been made (description of the scenes, lengths of cuts, sections of law that are referred to, and so on). There are 14 columns in this file. The number of record instances (rows) is 8330.

1.5.2 Spreadsheets from a government agency committee

This data set is about Swedish elections, and consists of three spreadsheets. They contain information about the candidates on different levels:

- Central level (central government)
- County/region level (county council)
- Municipal level (municipal executive board)

For the municipal level, there is also information about the electorate.

Each spreadsheet contains information about distribution with respect to sex, age, ethnic background, education level, and income.

1.5.3 Digitised documents (images of different formats)

See description for Scenario 2.2 (**DCH-RP_WP5_Scen-2-2_ID-66.doc**).

1.5.4 The web site “Linnéjubiléet” (Linné Anniversary)

See description for Scenario 2.4 (**DCH-RP_WP5_Scen-2-4_ID-68.pdf**).